Studies in Educational Evaluation 67 (2020) 100916

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment

Xiaoming Zhai[a,*], Kevin C. Haudek[b,c], Molly A.M. Stuhlsatz[d], Christopher Wilson[d]

[a] *Department of Mathematics and Science Education, University of Georgia, United States*
[b] *CREATE for STEM Institute, Michigan State University, United States*
[c] *Department of Biochemistry and Molecular Biology, Michigan State University, United States*
[d] *BSCS Science Learning, United States*

ARTICLE INFO

Keywords:
Construct-irrelevant variance
Machine learning
Pedagogical content knowledge
Science teacher
Constructed response assessment

ABSTRACT

Machine learning has been frequently employed to automatically score constructed response assessments. However, there is a lack of evidence of how this predictive scoring approach might be compromised by construct-irrelevant variance (CIV), which is a threat to test validity. In this study, we evaluated machine scores and human scores with regard to potential CIV. We developed two assessment tasks targeting science teacher pedagogical content knowledge (PCK); each task contains three video-based constructed response questions. 187 in-service science teachers watched the videos with each had a given classroom teaching scenario and then responded to the constructed-response items. Three human experts rated the responses and the human-consent scores were used to develop machine learning algorithms to predict ratings of the responses. Including the machine as another independent rater, along with the three human raters, we employed the many-facet Rasch measurement model to examine CIV due to three sources: variability of scenarios, rater severity, and rater sensitivity of the scenarios. Results indicate that variability of scenarios impacts teachers' performance, but the impact significantly depends on the construct of interest; for each assessment task, the machine is always the most severe rater, compared to the three human raters. However, the machine is less sensitive than the human raters to the task scenarios. This means the machine scoring is more consistent and stable across scenarios within each of the two tasks.

1. Introduction

Machine learning has great potential for the evaluation of constructed responses, which is supported by findings that the inter-rater reliability between human ratings and machine predicted ratings is comparable to that shown between two trained human coders (Zhai, Yin, Pellegrino, Haudek, & Shi, 2020; Liu et al., 2014a; Nehm, Ha, & Mayfield, 2012; Shermis, 2015; Zehner, Saelzer, & Goldhammer, 2016). The outcomes of machine learning-based assessments provide immediate feedback to teachers and students, and thus machine learning is increasingly used in web-based inquiry, game-based assessment, simulation assessment, and adaptive learning (Zhai, 2019; Zhai, Haudek, Shi, Nehm, & Urban-Lurain, 2020). More importantly, machine scoring shows little repetitive scoring bias (Clauser, Kane, & Swanson, 2002). That is, compared to human scorers, the machine can precisely assign the same score to the same response on multiple occasions, while human scorers might not always be able to accomplish this. Given these advantages of machine scoring, however, we noticed that prior studiesprimarily focused on examining how precisely the machine scores reflect the construct of interest, the latent trait of examinees that a test is intended to assess (e.g., domain knowledge and cognitive ability), which accounts for the variance of the examinees' performance in tests (Cronbach & Meehl, 1955). No data to date show evidence of how machine scoring reflects the construct-irrelevant variance (CIV, also called *error variance*), which may lead to score misinterpretation (Gallagher, Bennett, Cahalan, & Rock, 2002; Messick, 1984). CIV can arise due to psychological or situational factors such as rater severity or contextualized features of the assessment. Without examining the CIV in automatically scored assessments, we cannot understand whether using the results from machine scoring will increase or decrease the test validity, as compared to the results of human scoring.

This study fills the gap by comparing machine and human scoring of constructed responses with rich contextualized information in a teacher pedagogical content knowledge (PCK) assessment. We focus on two

* Corresponding author at: 105J Aderhold Hall, 110 Carlton Street, Athens, Georgia, United States.
E-mail address: xiaoming.zhai@uga.edu (X. Zhai).

https://doi.org/10.1016/j.stueduc.2020.100916
Received 15 September 2019; Received in revised form 29 July 2020; Accepted 31 July 2020

sub-constructs of PCK: the teacher's ability to a) analyze student thinking, and b) respond to student thinking with appropriate pedagogical moves. Both are regarded as fundamental in teachers’ competence to transform content knowledge into a form of knowledge that is pedagogically useful for teachers (Magnusson, Krajcik, & Borko, 1999; Shulman, 1986). Effective science teaching should be based on a thorough understanding of student understanding, which requires teachers to responsively adapt or adjust teaching strategies in order to improve student thinking. The study focuses on contextualized constructed-response assessment because this type of assessment is deemed more likely to yield CIV (Haladyna & Downing, 2004; Zaichkowsky, 1985). Contextualized open-ended items are popularly used in science assessments as they are more authentic and, consequently, examinees might be more engaged in the test and perform in a way that reflects their competency (Zhai et al., 2019). However, those rich contexts or scenarios and the approach of rating the responses might result in scores that do not entirely reflect the construct of interest, thus compromising the interpretation and use of the test scores. A prior study suggests three sources of CIV might be involved in the contextualized constructed-response assessment (Zhai, Haudek, Stuhlsatz, & Wilson, 2020): variability of the scenario, rater severity, and rater sensitivity of scenarios. In this study, we investigate whether those three sources of CIV are also present when response scores are assigned by a machine learning algorithm. We employ both human scorers and an automated scoring system to score teachers’ responses to a set of video-based assessments targeting two sub-constructs of PCK: Analyzing Student Thinking and Analyzing Responsive Teaching. We asked the following research questions:

(a) How much CIV is drawn by the variability of scenarios, the rater severity, as well as the judging sensitivity of the scenarios in the PCK assessment? What is the difference of the variance between the two sub-constructs of PCK?
(b) To what degree does the machine scoring approach differ from that of the human scorers with regards to the extent of CIV?

In the following sections, we first review the assessment toward science teachers’ PCK. Then, we introduce the machine scoring and human scoring of constructed responses. Finally, we present the empirical research and findings from the measure of a video-based PCK assessment.

2. Assessing science teachers’ pedagogical content knowledge

Within Shulman's landscape, teacher pedagogical content knowledge, the “ways of representing and formulating the subject that make it comprehensible to others [students]” (Shulman, 1986, p. 9), has been regarded as the most prominent element of science teachers' professional proficiency (Brookhart, 2011) and the knowledge unique to the profession of teachers (Magnusson et al., 1999; Shulman, 1986). A proficient science teacher should not only possess appropriate subject matter knowledge, but also the knowledge of how to guide and scaffold students’ effective learning (Kleickmann et al., 2013; Park & Oliver, 2008). This knowledge, which is visualized in teachers' pedagogical practice, can be specified as the extent to which teachers can anticipate students’ thinking of the subject matter knowledge on-the-fly, and the extent to which teachers can responsively adjust pedagogy to promote students' learning (Keller, Neumann, & Fischer, 2017; Wilson, Borowski, & van Driel, 2019). These two aspects have been regarded as being fundamental to teachers’ PCK—transforming content knowledge into a form that is pedagogically useful (Alonzo, Kobarg, & Seidel, 2012). However, both of these performance-based constructs are not easy to measure using traditional paper and pencil assessments because teachers may not be aware that they are engaging in PCK and they do not always possess the language to express their PCK (Kagan, 1990). In contrast, performance-based measures are more capable of eliciting teachers' PCK but are regarded as both time-consuming to administrate and challenging to grade due to the diversity of teaching between teachers (Baxter & Lederman, 1999).

Recently, using video clips as part of an assessment has been recommended as an efficient approach to measuring teachers' professional development and is posited to cover the shortcomings mentioned above (Alonzo & Kim, 2016; Zhang, Lundeberg, Koehler, & Eberhardt, 2011). Compared to using only paper and pencil instruments, asking teachers to respond to questions after observing a given teaching video may better elicit teachers’ PCK. This is because, PCK is performance-based in nature and is expected to be elicited when teachers are in authentic teaching scenarios. In this sense, observing teaching video clips has an advantage over surveying teachers with paper and pencil questions, as the latter has limited capacity to engage teachers in authentic teaching scenarios. Also, watching deliberately selected video clips might significantly reduce the testing time while retaining a nearly equal capability of capturing the richness and complexity of elusive classroom teaching practices (Santagata, 2009). More importantly, by using video clips, teachers’ scores can be easily standardized and compared, thus increasing the test reliability.

However, video-based PCK instruments are also subject to test validity issues, such as irrelevant constructs as recognized in a prior study, in which Zhai, Haudek, Stuhlsatz, and Wilson identified three critical sources that may yield CIV in video-based PCK measures.

3. Constructed-irrelevant-variance in assessing teachers’ PCK

A construct is the latent trait of examinees that a test is intended to assess (e.g., teachers' PCK), accounting for the variance of examinees' performance on tests (Cronbach & Meehl, 1955). Haladyna and Downing (2004) suggest that identifying a construct of interest is the base of designing assessment tasks. A valid conclusion based on the assessment scores relies on the consistency between assessment scores and the construct of interest. However, as Messick (1984) suggested, irrelevant constructs might cause variances in the scores, which will contaminate the inference and interpretation of the scores. For example, in a test targeting students' understanding of force and motion, a testing task with a heavy reading load might affects the variance of students' scores and might not lead to a valid inference of students' understanding of force and motion, the construct of interest. Instead, the variance of the scores reflects the construct that is not expected, such as reading speed in the example above. This part of the variance is regarded as CIV. Many studies suggest examining CIV in order to ensure test validity (AERA & NCME, 1999). Prior studies have suggested many sources that might cause CIV, such as test preparation, test development and administration, scoring, examinee preferences and cheating (Gallagher et al., 2002; Haladyna & Downing, 2004). In this study, we used a video-based measure to assess teachers’ PCK. Assessment questions were developed based on the scenarios and then graded by multiple experts. According to a model for CIV (Zhai, Haudek, Stuhlsatz, & Wilson, 2020) that specifically focuses on contextualized constructed response assessments, three sources of CIV might be prominent on contaminating the testing validity: variability of scenarios, rater severity, and rater sensitivity of scenarios. The below section explains the definitions and rationale of the three sources of CIV.

3.1. *Variability of scenarios*

The scenario is broadly defined by the science storyline occurring in the teaching video, which provides the basis for observant teachers’ analysis of teaching and learning. The major contributors to scenarios include the context, the topic of teaching, and the characters (e.g., teacher and students). Assuming we administer a question multiple times and each time with a different scenario, it is expected that the observant teachers’ performance will have the least variance due to the scenario variability or the performance differences between teachers. If not, for example, a teacher overperforms on one scenario than on another scenario, the testing score could not support a valid inference about the teacher's ability. It should be noted that if this variability is random there might not be much harm to the validity—increasing the number of testing tasks can help to reduce the variability. However, if this variance is systematic, the scores assigned to teachers' performance cannot support a valid conclusion (e.g., the teacher is competent or not with regard to PCK). Therefore, in order to support the valid inference, the variability of scores is expected to be due only to the construct of interest (i.e., teachers' PCK that the question is designed to assess). However, an observant teachers' performance might be inevitably dependent on scenarios because teachers might experience varied challenges to interpreting or understanding the scenarios before answering the question; this would consequently cause CIV. In this sense, it is critical to examine and control the variability of the scenarios that might cause CIV.

3.2. Rater severity

We typically employ multiple raters to score teachers' constructed responses (i.e., performance). However, raters might exhibit different rater severity even using the same scoring criteria. Some raters tend to give the same response higher scores than other raters do. That is, the observant teachers' scores might vary due to the severity of the rater(s) rather than teachers' PCK. In this sense, there is CIV due to the rater severity contaminating the scores (Ebel & Frisbie, 1986; Goodwin, 2016; Ooi & Engelhard, 2019).

3.3. Rater sensitivity of scenarios

While the variation of scenario is a feature of the assessment and the rater severity is a feature of the individual rater, Zhai, Haudek, Stuhlsatz, and Wilson (2020) realized an interaction effect between the two factors. That is, even the same rater may have varied severity under different assessment scenarios because of the variation of scenario features, such as richness and complexity. These scenario features might impact raters' interpretation of the observant teachers' performances (i.e., responses) thus causing systematic bias to teachers' scores. Therefore, examining and controlling for the rater sensitivity about the scenarios is critical to ensure test validity.

4. Machine scoring vs. human scoring on constructed responses

Machine learning has been widely applied in the evaluation of constructed responses and has demonstrated comparable inter-rater reliability with human raters (Zhai, Yin, Pellegrino, Haudek, & Shi, 2020; Liu et al., 2014b; Shermis, 2015). Using a developed machine learning algorithm could significantly reduce the grading time spent on constructed responses and provide teachers with timely feedback to help teachers adjust teaching. The core of applying machine learning in the automated scoring of assessments is to first develop and test classification algorithms; after which, these scoring algorithms can be applied to new sets of responses in order to predict scores or codes. Different from traditional computerized assessments by which computers primarily assign scores according to instructions set by humans, the machine 'learns' from a set of human-labeled data and adjusts algorithmic parameters from what it learns; thus, such an approach is called supervised machine learning(Zhai, Krajcik, & Pellegrino, 2020). After a validation phase, the algorithm(s) can be applied to new data and predict the labels of a new set of data.

Prior studies have collected evidence that the machine-human interrater reliability and the accuracy of the scores are sufficient to be used both in large-scale and classroom assessments (e.g., Zhai, Yin, Pellegrino, Haudek, & Shi, 2020), but these results did not address the CIV in the machine predicted scores. Without this information, we don't know to what extent the machine scores are biased by different sources that account for the CIV, even if it sufficiently meets performance metrics (e.g., accuracy, precision or inter-rater reliability with humans). Since machines 'learn' from the human-labeled data as part of the training process of machine learning models, we might expect that these scoring algorithms would generate scores with CIV that is 'similar' to human scorers. For example, as it has been shown that humans tend to apply a more severe criterion when assigning scores to a task with a given scenario as compared to that with other scenarios, we don't know whether the machine scores are similarly influenced by a given assessment scenario. No data indicate how similar the machine performs as compared to human scorers with regards to CIV. Given that CIV has drawn attention to computerized testing since the turn of the 21st century (AERA & NCME, 1999), we argue it is necessary to examine and compare the machine and human scores with regard to CIV. This information can help us to better design assessment tasks that can be automatically scored by computers.

In this study, teachers' short responses to the assessment tasks are first rated by three trained human experts. After a consensus procedure, the consensus scores are used to train machine learning algorithms to develop a scoring model. As part of this process, responses that are already scored by humans (i.e., the training set) also receive a "predicted" score from the scoring model. Afterward, the scores assigned to each response by humans and machine are compared. The focus of the study is to investigate how the machine-predicted score is different than human assigned scores with regards to CIV drawn by the variability of scenarios, the rater severity, as well as the judging sensitivity of the scenarios.

5. Methods

5.1. Sample

We recruited teacher participants from the United States for this study through formal email invitations. To ensure that our sample of respondents included teachers with the requisite teaching knowledge to perform highly on the assessment, we targeted in-service US teachers who had recently participated in a PCK-focused teacher professional development program, received a presidential award in science, or were National Education Association certified. To be included in the study, participants had to be currently employed as a teacher and teaching science in grades 3 through 12. Using these criteria, 187 teachers qualified to participate in the study. 12.3 % of the sample taught grades 3–5, 49.3 % taught grades 6–8, and 38.5 % taught grades 9 − 10. The sample had an average of 16 years of teaching experience and 77.5 % identified as female.

5.2. Instrument

For this study, we used an online video-based analysis approach to assess teachers' ability to analyze two distinct sub-constructs of PCK, identifying student thinking and identifying teacher's responsive teaching (Kirschner, Taylor, Rollnick, Borowski, & Mavhunga, 2015). First, we developed a pool of science teaching video clips ranging from 3 − 5 minutes. These video clips were identified from full-length science lessons recorded in elementary school classrooms in the United States. The initial video selection process was intended to target video clips that were most likely to elicit the sub-constructs of interest. Four experts with comprehensive knowledge of the sub-constructs of interest and content knowledge in elementary science viewed the pool of videos and then came together to agree on the most appropriate video clips. Eventually, the group of experts selected eleven video clips according to the content analysis. Based on which, the experts confirmed the content validity.

We then piloted the 11 video clips with 192 science teachers (different from the 187 science teachers in the formal test) to test the extent to which each of the eleven videos elicits the sub-constructs. We

Table 1
Examples of constructed responses and the scores.

Scenario introduction	Score	Analyzing student thinking	Analyzing responsive teaching
S1. This 4th-grade class is learning about how airplanes fly. In the first part of this lesson, they discuss the concept of "drag" and what it means. Please view the video and respond in complete sentences to the two questions below the video.	0	The students do not have a great understanding of the concepts. All of the students are not actively involved in formulating explanations. Students do not seem very confident in their ideas as they are related to the science content.	The teacher does not acknowledge when students' ideas/understandings are accurate or good examples. Many great examples were given, but not much positive feedback. The focus was sometimes more on certain words/meaning rather than the student's ability to use the word correctly.
	1	The students seem to have an idea about what drag is but are having trouble really explaining what it is or how it works in real-world situations. They are able to come up with related examples that involve drag and are connecting it to the idea of resistance, but it is hard to tell if they understand the forces operating that create a drag situation.	I notice the teacher pushing students to think about drag more broadly. She helps to connect examples students give to other ideas already shared. She pushes students to explain their ideas more fully (i.e., what does resistance mean?) She then discusses resistance by bringing in an experience that most students have had and uses that to help students think about resistance.
S2. In this 5th grade class, students are trying to explain different scenarios that were provided to their groups (e.g., condensation on the outside of a cold soda can). The teacher instructs the groups to explain their respective scenarios in terms of evaporation, condensation, and water molecules. Please view the video and respond in complete sentences to the two questions below the video.	0	Students have different ideas and are connecting to background knowledge and experiences from their lives. Some ideas are closer than others. Students do start to clarify their thinking as they continue.	Yikes! I tried to follow the teacher in this lesson about condensation, but she left many ideas open for the students to translate themselves.
	1	The students have a variety of ideas related to water molecules and how condensation occurs. The students also make interesting connections to other similar ideas like that of a person sweating. However, some of the student ideas are incomplete or not completely correct.	The teacher continues to ask students questions to push them towards the idea that it is when the water vapor cools next to the can, the water condenses forming droplets. She tries first to get students to think about where and when the air will cool down. She often tells her students they are on the right track and then asks additional questions to try to correct misconceptions.
S3. This 5th-grade class is on the fifth of six lessons in a unit on the water cycle. Working in groups, students have completed an investigation using a distillation apparatus to follow the evaporation and condensation of water. The groups are working together to explain the observations they made in this investigation.	0	Students seem to have a high degree of comfort talking about the science content in the video. They are able to explain and respond to teacher questions using the diagram.	The teacher asks students to explain using their diagrams. The teacher asks students to explain using their experience/ observations from doing the experiment. The teacher asks them to explain their ideas using a molecular view.
	1	The students did a great job of using the experience of the lab as well as the diagram to help explain their thinking. They still struggled to get to the point of how the molecules were moving (slowing down and getting closer together.)	The teacher asks students to explain using their diagrams. The teacher asks students to explain using their experience/ observations from doing the experiment. The teacher asks them to explain their ideas using a molecular view.

investigated the percentage of pilot teacher responses (by video) that led to positive scores in each sub-construct (Supplementary Table S1), noting that the frequency by video scenario differed. Using the frequency information, we chose three videos (S1, S2, and S3) to include in the final data collection. Based on the pilot testing data, these three videos were the most likely to elicit teachers' PCK regarding the two sub-constructs and avoid floor effects. These videos included a fourth-grade classroom learning about air resistance (video scenario 1), a fifth-grade classroom learning about condensation on the outside of soda can (video scenario 2), and a fifth-grade classroom learning about distillation (video scenario 3). A description of the three video clips is provided in Table 1.

For the data collection reported here, participating teachers were first asked to view an individual video clip (scenario). After viewing each scenario, teachers were asked to respond to two prompts designed to measure their ability to analyze student thinking (Task I) from the video and connect student thinking to how the teacher in the video responds to student thinking (Task II). The first task prompt asks "What do you notice about the student ideas related to the science content in this video?" and the second task prompt asks, "What do you notice about how the teacher responds to student ideas related to the science content in this video?" Respondents were provided with a text box to record their written responses to each prompt. For each of the three videos, each participant viewed a video scenario, then answered the two prompts. This resulted in a total of six written responses for each participant; three for Analyzing Student Thinking and three for Analyzing Responsive Teaching.

5.3. Data collection

We used the Qualtrics survey platform to collect data for this study. First, teachers were asked to respond to a series of questions that determined whether they met the inclusion criteria for the study. Teachers were assigned to view each of the three video clip scenarios and respond to the two prompts in random order. After viewing each clip and responding to the prompts they were asked to respond to a demographic questionnaire. The instrument was completed in 30 − 40 min.

5.4. Scoring

5.4.1. Human scoring

We developed a binary coding framework to score teachers' constructed responses. For Task I, the rubric highlights whether teachers can successfully recognize students' thinking based on the evidence they observed from the video; for Task II, the rubric focuses on whether teachers can effectively identify how the teacher in the video applied pedagogical strategies in order to respond to and improve student thinking. Coding examples are presented in Table 1.

Three human raters, given the pseudonyms, Lacie, Tony, and Emerson, participated in a training routine to establish an inter-rater agreement. The raters were chosen based on their substantial expertise in the theoretical components of teacher PCK, elementary science content knowledge, and prior scoring experience. After completing this process, each rater received an equal number of random responses to score, with each response scored by at least two of the raters. Meantime, the raters were blinded to the grade level of the participants and other demographic factors. Interrater reliability was sufficient for Analyzing Student Thinking (Task I), but not for Analyzing Responsive

Table 2
Human-human agreement and Human-machine agreement of the scoring.

Construct	Human-human	Human-machine
	Cohen's κ	Cohen's κ
Task I	κ_{te} = .869, κ_{tl} = .882, κ_{el} = .878	.805
Task II	κ_{te} = .520, κ_{tl} = .428, κ_{el} = .496	.570

Teaching (Task II). Eventually, discrepancies in human scores were solved by consensus discussion between the three raters, until everyone agreed with the scores (see Table 2). Task II had a lower interrater agreement than Task I (Table 2), which suggests that Task II might be more challenging for raters to score and agree upon than Task I. This might be due to the fact that teachers had multiple choices to respond to student thinking, which made the responsive teaching more complex for raters to meet an agreement with the scores assigned. The vast range of interrater reliability (κ = .428–.882) between the two tasks provided us opportunities to explore the computer's performance patterns with both easy and challenging tasks for humans to grade.

5.4.2. Machine scoring

We developed a machine learning application, which uses a text classification approach to assign a score to each teacher response (see Aggarwal & Zhai, 2012). To generate a single data set for each task, we combined all teacher responses across the three scenarios into a single data set. This set of human-scored responses with consensus scores was used as a training set to develop a scoring algorithm for each task, using an ensemble approach of eight different machine learning classification algorithms by AACR web app (AACR, 2019). Text from the responses in the training set was extracted and a lexical feature set was generated. These features were used as independent variables in the machine learning algorithms to predict human assigned codes or scores. The output predictions from each of the eight algorithms were combined into a single output prediction, which was taken as the machine-assigned score. The eight outputs were combined using a cross-validated weighted voting scheme, in which the overall accuracy of each algorithm was used to weight its prediction for a given response in the overall machine-assigned score output for that response (Large, Lines, & Bagnall, 2019). The computer scoring model was generated using a 10-fold cross-validation approach with the training set of data (Nehm et al., 2012). The performance of the scoring model can be measured by comparing human-assigned scores to computer-assigned scores within the training set of data.

5.5. Analysis

According to Zhai, Haudek, Stuhlsatz, and Wilson (2020), if the examinee scores are under constraints by interrelated conditions, the data would experience local dependency, and, thus, the classical Rasch model is not applicable. In this situation, the many-facet Rasch model can be applied to the data without concern for the local independence assumption (Linacre, 1989). In our study, we applied a four-facet (i.e., teacher ability, the variability of scenarios, rater severity, and the judging sensitivity to the specific scenario) Rasch model to each testing task. The formula is,

$$log\left(\frac{P_{nijk}}{P_{nijk-1}}\right) = B_n - D_i - F_k - C_j - S_{ij}$$

Where

P_{nijk} represents the probability of examinee n is awarded a score of k on item with scenario i by the rater j; P_{nijk-1} indicates that the probability of examinee (teacher) n is awarded a score of k-1 on the item with scenario i by rater j; B_n is the ability measure for examinee n; D_i is the difficulty measure of the item with scenario i, F_k is the difficulty measure of stepping from rating level k-1 to k. The C_j represents the rater severity of rater j, and the S_{ij} represents the rater j's scoring sensitivity to the item with scenario i.

Using this many-facet Rasch measurement model, all the measures of the four facets share a logit scale so that we could compare the parameters within and between facets. First, we could compare the rater severity measure between the machine and the human raters (within a facet) by a Chi-square test. The Chi-square test will accompany the separation reliability to test for significance. For the separation reliability,

$$R = \frac{s - \sum_{i=1}^{T} \tau_i}{s}$$

Where s is the variance of the parameter estimates,

$$s = \frac{1}{T-1}\sum_{i=1}^{T} (\delta_i - \delta_0)^2$$

and the Chi-squared value is,

$$X^2 = \sum_{i=1}^{T} \frac{\delta_i^2}{\tau_i}$$

Where, δ_i is the estimated parameter in the facet, δ_0 is the mean of the estimated parameters, τ_i is the estimated error variance, and T is the total number of the parameters. According to the definition, the separation of reliability is a measure ranging from 0 to 1. A value close to 0 indicates the parameters are homogeneous, while a value close to 1 indicates the parameters are heterogeneous. The larger the value of the Chi-square, the more heterogeneous the parameters, and therefore the greater the CIV.

We could also compare the measures of machine severity and machine sensitivity of the scenario (between facets). Since the focus of the study is to examine CIVs generated due to the variability of item difficulty with different scenarios, the variability of rater severity, and the variability of the judging sensitivity of the scenarios, we centered each of the three facets at zero logit. Therefore, the variabilities of the measures between facets are comparable. By calibrating the variance of each facet, we could estimate the contributions of these facets to the CIVs. The greater the variance within a facet, the more CIV is yielded due to this facet. For example, if the variance of the rater severity is greater, then the variance of scores will be greater, not due to the underlying construct of interest, but instead due to the variability of the rater severity.

6. Result

To answer our research questions, we first examine the statistics of the measurement model and then present the CIVs for the two tasks, respectively. Then, we compare the machine performance and human rater performance across the two tasks.

6.1. Statistics of the measurement model

In the study, there are three scenario-based questions in each task and each response is rated by four raters (i.e., three humans and one computer). Thus, each participating teacher received a total of 12 scores for each testing task. We took each of these scores as an item and thus examined the quality of the 12 items for Task I and those for Task II respectively through both the item response test (IRT) model fit and the classical statistical model fit (See Table 3).

6.1.1. IRT model

We first examined the mean square residuals (MNSQ) between the observed and predicted scores using the sum of the Chi-square residuals and the degrees of freedom (Wu, Adams, Wilson, & Haldane, 2007). The Outfit MNSQ is the unweighted mean square residual between the observed score and the predicted score, while the Infit MNSQ is the weighted mean square residuals calibrated by assigning more weight to

Table 3
Item response model fit and traditional statistical model fit.

Rater /scenario		Item response test model fit						Classical statistical model fit					
		Outfit MNSQ	CI	T	Infit MNSQ	CI	T	Score	Count (%)	Pt Bis	t (p)	Person ability	Disc
Tony	S1	0.97	(0.76, 1.24)	−0.2	1.16	(0.78, 1.22)	1.4	0	56 (43)	−0.65	−9.62 (.000)	−0.94	0.65
								1	74 (57)	0.65	9.62 (.000)	1.84	
	S2	0.86	(0.76, 1.24)	−1.2	1.04	(0.78, 1.22)	0.4	0	52 (40)	−0.69	−10.97 (.000)	−1.06	0.69
								1	79 (60)	0.69	10.97 (.000)	1.74	
	S3	0.76	(0.75, 1.25)	−2.1	0.98	(0.79, 1.21)	−0.2	0	56 (44)	−0.71	−11.13 (.000)	−1.04	0.71
								1	71 (56)	0.71	11.13 (.000)	1.93	
Emerson	S1	0.86	(0.77, 1.23)	−1.1	1.03	(0.77, 1.23)	0.3	0	54 (39)	−0.72	−12.20(.000)	−1.27	0.72
								1	86 (61)	0.72	12.20 (.000)	1.87	
	S2	0.67	(0.76, 1.24)	−3.0	0.92	(0.78, 1.22)	−0.7	0	56 (42)	−0.77	−13.76 (.000)	−1.27	0.77
								1	76 (58)	0.77	13.76 (.000)	1.96	
	S3	0.62	(0.75, 1.25)	−3.4	0.86	(0.78, 1.22)	−1.3	0	59 (47)	−0.77	−13.59 (.000)	−1.06	0.77
								1	66 (53)	0.77	13.59 (.000)	2.12	
Lacie	S1	0.79	(0.75, 1.25)	−1.7	0.98	(0.75, 1.25)	−0.1	0	53 (42)	−0.77	−13.56 (.000)	−1.39	0.77
								1	72 (58)	0.77	13.56 (.000)	2.40	
	S2	0.65	(0.75, 1.25)	−3.2	0.98	(0.73, 1.27)	−0.1	0	43 (34)	−0.76	−12.92 (.000)	−1.48	0.76
								1	83 (66)	0.76	12.92 (.000)	1.94	
	S3	0.63	(0.75, 1.25)	−3.4	0.9	(0.76, 1.24)	−0.8	0	55 (44)	−0.80	−14.57 (.000)	−1.22	0.80
								1	69 (56)	0.80	14.57 (.000)	2.35	
Machine	S1	1.01	(0.80, 1.20)	0.1	1.10	(0.81, 1.19)	1.0	0	78 (41)	−0.69	−13.11 (.000)	−1.16	0.69
								1	112 (59)	0.69	13.11 (.000)	1.97	
	S2	0.85	(0.80, 1.20)	−1.5	1.00	(0.81, 1.19)	0.1	0	80 (42)	−0.76	−15.94 (.000)	−1.16	0.76
								1	111 (58)	0.76	15.94 (.000)	1.99	
	S3	1.03	(0.80, 1.20)	0.3	0.93	(0.82, 1.18)	−0.8	0	86 (46)	−0.70	−13.43 (.000)	−1.03	0.70
								1	101 (54)	0.70	13.43 (.000)	2.11	

Rater/ scenario		Item response model fit						Traditional statistic model fit					
		Outfit MNSQ	CI	T	Outfit MNSQ	CI	T	Score	Count (%)	Pt Bis	t (p)	Person ability	Disc
Tony	S1	0.84	(0.76, 1.24)	−1.3	0.95	(0.81, 1.19)	−0.5	0	80 (62)	−0.63	−8.42 (.000)	−1.39	0.60
								1	49 (38)	0.63	8.42 (.000)	0.27	
	S2	0.94	(0.76, 1.24)	−0.4	1.02	(0.81, 1.19)	0.3	0	81 (62)	−0.63	−8.17 (.000)	−1.24	0.59
								1	49 (38)	0.63	8.17 (.000)	0.14	
	S3	0.81	(0.75, 1.25)	−1.6	0.93	(0.82, 1.18)	−0.7	0	57 (45)	−0.59	−7.88 (.000)	−1.60	0.58
								1	70 (55)	0.59	7.88 (.000)	−0.08	
Emerson	S1	0.91	(0.76, 1.25)	−0.7	0.99	(0.82, 1.18)	−0.1	0	74 (58)	−0.58	−8.01 (.000)	−1.56	0.58
								1	54 (42)	0.58	8.01 (.000)	−0.04	
	S2	0.84	(0.76, 1.25)	−1.3	0.92	(0.76, 1.24)	−0.7	0	98 (77)	−0.64	−8.26 (.000)	−1.34	0.59
								1	30 (23)	0.64	8.26 (.000)	0.44	
	S3	0.65	(0.75, 1.25)	−3.1	0.84	(0.79, 1.21)	−1.6	0	86 (69)	−0.65	−9.01 (.000)	−1.52	0.63
								1	39 (31)	0.65	9.01 (.000)	0.38	
Lacie	S1	0.80	(0.75, 1.25)	−1.7	0.91	(0.82, 1.18)	−0.9	0	72 (58)	−0.66	−8.97 (.000)	−1.43	0.63
								1	53 (42)	0.66	8.97 (.000)	0.14	
	S2	1.11	(0.75, 1.25)	0.9	1.06	(0.82, 1.18)	0.7	0	65 (52)	−0.63	−8.19 (.000)	−1.55	0.59
								1	60 (48)	0.63	8.19 (.000)	0.08	
	S3	0.70	(0.77, 1.23)	−2.8	0.85	(0.83, 1.17)	−1.8	0	94 (63)	−0.68	−10.94 (.000)	−1.51	0.67
								1	54 (36)	0.68	10.94 (.000)	0.41	
Machine	S1	0.84	(0.80, 1.20)	−1.6	0.96	(0.82, 1.18)	−0.4	0	137 (72)	−0.61	−10.45 (.000)	−1.29	0.61
								1	53 (28)	0.61	10.45 (.000)	0.40	
	S2	1.04	(0.80, 1.20)	0.4	1.07	(0.83, 1.17)	0.8	0	136 (71)	−0.52	−8.45 (.000)	−1.26	0.52
								1	56 (29)	0.52	8.45 (.000)	0.30	
	S3	0.80	(0.80, 1.20)	−2.1	0.97	(0.82, 1.18)	−0.3	0	133 (71)	−0.59	−10.05 (.000)	−1.30	0.59
								1	54 (29)	0.59	10.05 (.000)	0.38	

the responses that fit the model (Liu & McKeough, 2005). The prior is more sensitive to items with a difficulty close to the examinee ability, while the latter is more sensitive to extreme cases, such as the extremely difficult or easy items. Both measures are suggested by Linacre (2002) to be productive within the range of 0.5 − 1.5. In our study, both the Infit (ranging from 0.62 to 1.11) and Outfit (ranging from 0.85 to 1.16) MNSQ for all 24 items across the two tasks are within the productive range (see Table 3). The 95% confidential interval is also suggested as useful criteria for model fit. While the Infit MNSQ for all of the 24 items are within the interval, the Outfit MNSQ for four items in Task I and two items in Task II are outside of the 95% confidential intervals, which indicates these items are less productive than the other items. This less than desired fit is also indicated by the T value, whose absolute value is greater than 2 for these items.

6.1.2. *Classical statistical model*

Compared to the IRT model fit, the classical statistical model fit highlights the nature of the raw score. First, we examined the teachers' distribution on the binary scale (i.e., level 0 and 1) and found nearly every level is distributed with more than 30 % teachers, with a single exception of 28 % for level 1 of Scenario 3 of Task II, as rated by the machine. The point-biserial correlation (i.e., discrimination) that is calibrated between the categories of scores and the examinee's total score are all greater than .60 (absolute value) with only a few exceptions. Also, the teachers' average ability monotonically increases with the levels of response for all 24 items. These measures suggest a robust item discrimination.

Table 4
Variance and separation measure of the four facets for Tasks I and II.

	Task I				Task II			
Parameter #	Teacher	Scenario	Rater severity	Rater sensitivity	Teacher	Scenario	Rater severity	Rater sensitivity
1	–	−0.078 (0.088)	−0.028 (0.095)	0.110 (0.117)	–	−0.068 (0.073)	−0.373 (0.082)	0.373 (0.100)
2	–	−0.219 (0.088)	0.025 (0.096)	0.019 (0.116)	–	0.146 (0.074)	0.231 (0.084)	0.223 (0.100)
3	–	0.297 (0.124)	−0.074 (0.098)	−0.129 (0.165)	–	−0.078 (0.104)	−0.345 (0.081)	−0.596 (0.141)
4	–	–	0.077 (0.166)	−0.200 (0.117)	–	–	0.487 (0.142)	−0.557 (0.102)
5	–	–	–	0.147 (0.117)	–	–	–	0.434 (0.104)
6	–	–	–	0.052 (0.166)	–	–	–	0.123 (0.145)
7	–	–	–	0.197 (0.120)	–	–	–	0.063 (0.099)
8	–	–	–	−0.301 (0.119)	–	–	–	−0.478 (0.098)
9	–	–	–	0.104 (0.169)	–	–	–	0.414 (0.140)
10	–	–	–	−0.107 (0.204)	–	–	–	0.120 (0.174)
11	–	–	–	0.135 (0.204)	–	–	–	−0.179 (0.174)
12	–	–	–	−0.028 (0.289)	–	–	–	0.059 (0.246)
Variance measure								
Variance	5.132	0.071	0.004	0.024	1.640	0.016	0.183	0.137
Mean	0.598	0	0	0	−0.776	0	0	0
N	203	3	4	12	203	3	4	12
Separation measure								
Separation reliability	–	0.235	0	0.659	–	0.765	0.942	0.945
χ^2 test for parameter equality	–	7.06	0.74	14.48	–	4.79	46.62	90.24
Degrees of freedoms	–	2	3	6	–	2	3	6
p-value	–	.029	0.865	.025	–	.091	.000	.000

Note. The numbers (i.e., 1–9) under Values for the *Scenario*Rater* facet correspond to: 1= S1*Tony, 2= S2*Tony, 3= S3*Tony, 4= S1*Emerson, ...12= S3*Machine.

6.2. Construct-irrelevant variance within the two tasks

Table 4 shows the values for the four facets separated by task, the variance of the elements within each facet, and the separation measures for the elements within each facet.

Task I. We found the teachers' average ability measure is 0.598, slightly greater than the average item difficulty, which is fixed to 0 logit. This indicates that the items are appropriate in terms of the difficulty of testing the teachers. The variance of the teacher ability reflects the construct of interest. Compared to the variance within other facets, the variance = 5.132 suggests that the teacher ability primarily accounts for the probability of correct responses.

The variance of the components within the facets, except the examinee facet, indicates to what degree the facets contribute to the CIVs. With regard to Task I, the components of the scenario yield the greatest variance, 0.071. The separation reliability, indicating to what degree the elements in a facet are separated with each other (Wright & Stone, 1979), is relatively low, 0.235, for the item difficulty due to variability of scenario, which means the difference between components of the scenario facet is relatively small; but the Chi-square test suggests that the difference is statistically significant, $\chi^2(2,\ N = 3) = 7.06,\ p = .\ 029$. The finding suggests that the variability of scenarios substantially contributes to the CIV for this task.

The variance of the rater severity (0.004) is only 5.6 % of those for the scenario facet. At the same time, we found the separation reliability is 0 for the rater severity, and the Chi-square test further confirms that there is no significant difference between the four raters, $\chi^2(3,\ N = 12) = 0.74,\ p = .865$. This result suggests that the contribution of rater severity to the CIV is very limited.

In contrast, the variance of the rater sensitivity for the scenarios is 0.024 and the separation reliability for the components of the rater sensitivity is up to 0.659. This difference between the components, according to the Chi-square test, is significant, $\chi^2(6,\ N = 12) = 14.48,\ p = .\ 025$, suggesting that the contribution of the rater sensitivity for scenarios is substantial.

Task II. With regard to task II, the variance of the teacher ability is only one-third of that for Task I, 1.640. This might be because the questions are too difficult, as evidenced that the teachers' average ability, -0.776, is lower than the average item difficulty, 0. Compared to Task I, while the variance of the teacher ability decreases in Task II, we found the total variance within the other three facets increases. This indicates that for Task II, the accountability for the probability of correct responses increases for the CIVs and decrease for the construct of interest.

We found the variance of scenario components is identical to Task I (0.071), which might be due to the two tasks sharing the same three scenarios. The separation reliability is 0.765, but the differences between the components are not significant, $\chi^2(2,\ N = 3) = 4.79,\ p = .\ 091$, indicating the three scenarios for Task II do not make a significant difference in terms of difficulty.

However, the variance of the rater severity (0.183) is different from that found for Task I, and is more than twice the amount of the scenario facet. The separation reliability is .942. Chi-square test further confirmed that the differences between raters are significant, $\chi^2(6,\ N = 12) = 46.62,\ p = .\ 000$. This result indicates that the variability of rater severity substantially contributes to the CIV.

The variance of the rater sensitivity to scenarios (0.137) is almost twice the amount of the scenario facet. The separation reliability (.945) is even higher than the rater severity and the differences between components of the facet are also significant, $\chi^2(6,\ N = 12) = 90.24,\ p = .\ 000$. These findings also suggest a substantial contribution of rater sensitivity to the CIV.

6.3. Comparison of the construct-irrelevant variance between machine scores and human scores

One of the purposes of the study is to compare the CIVs yielded by machine and human experts. We had three trained human experts and the machine rate each of the six items across the two testing tasks. The discrepancy between machine and human expert assigned scores is reflected in two facets: rater severity and the rater sensitivity of the scenarios.

6.3.1. Rater severity

For the two tasks, we found the three human experts and the machine show similar patterns in terms of severity for the two tasks (Fig. 1), even though the range of severity is vastly different (Task I= −0.074 to 0.077; Task II= −0.373 to 0.487). That is, Lacie seems like

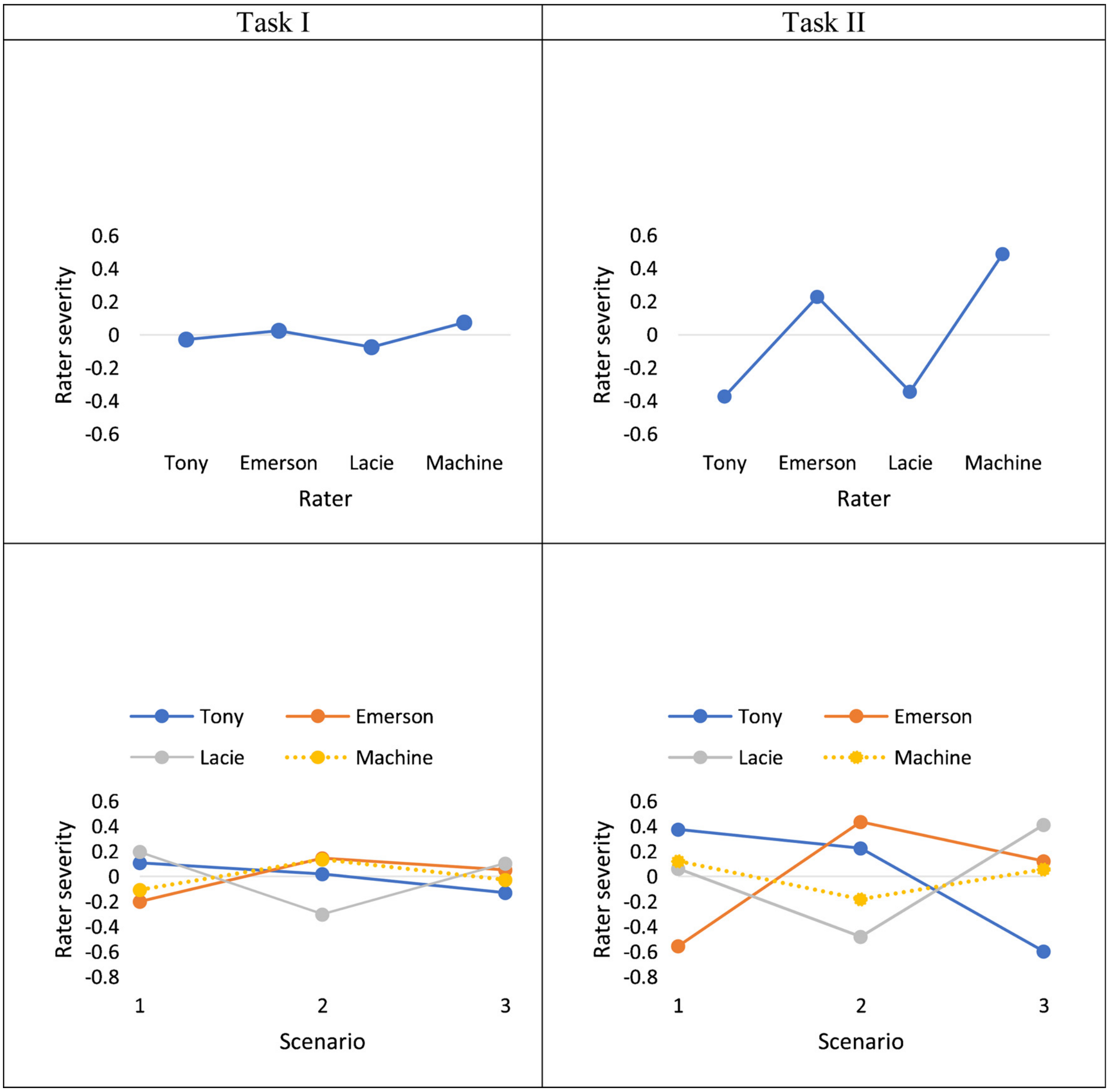

Fig. 1. The plot of rater severity and rater sensitivity of scenarios.

the most lenient rater, followed by Tony and Emerson, while the machine is the most severe rater for both tasks. At the same time, we also observe that the machine seems to hold the highest criteria to assign scores to teachers.

6.3.2. *Rater sensitivity of the scenarios*

Similar to the Rater severity facet, the Rater sensitivity facet also shows a significant range of sensitivity for the two tasks (Task I= −0.301 to 0.197; Task II= −0.596 to 0.434). By comparing the severity of the three human experts and the machine across the two tasks (Fig. 1), we found that the machine seems the most stable rater in terms of severity across item scenarios. This pattern is more distinct for Task II, in which the machine seems the least sensitive to the three scenarios compared to the human raters. Also, we note that human raters' patterns regarding sensitivity is different from each other. For example, the rater Emerson holds a severe criterion to assign scores for Scenario 1 (−0.557), a lenient criterion for Scenario 2 (0.434), and a medium criterion for Scenario 3 (0.123); while Lacie's pattern is almost the inverse of Emerson's (Scenario 1 = 0.063; Scenario 2= −0.478; Scenario 3 = 0.414) and Tony's pattern is different from both Emerson's and Lacie's. This variability leads to severe CIV.

7. Conclusion and discussion

Teacher PCK is a salient component of science teachers' professional proficiency and plays an important role in teacher professional development, as well as impacting students' achievement (Baxter & Lederman, 1999; Gess-Newsome et al., 2019; Kirschner, Borowski, Fischer, Gess-Newsome, & von Aufschnaiter, 2016; McNeill, González-Howard, Katsh-Singer, & Loper, 2016). However, assessing teachers' PCK is challenging because teachers may be unaware of their own PCK and may not always possess the language to express PCK. A performance assessment, which is more reliable to elicit and infer teachers' PCK, is regarded as both time-consuming and difficult to standardize scores. To address this issue, in our project, we developed a video-based PCK assessment. We first ask teachers to observe selected video clips of teaching with rich pedagogical information and then answer questions with regard to Analyzing Student Thinking and Analyzing Responsive Teaching in a short response format. Given that grading constructed responses is time-consuming, we developed machine learning scoring models to automatically score the responses. This study, which is within the context of a larger project to develop these PCK assessment tasks and automated scoring models, focused on examining the CIV yielded in the machine scores and compared it with that yielded in human scores. This study addresses the call of monitoring and uncovering the scoring bias which might affect test validity (Bejar, 2012) and helps us better understand the appropriateness of using a machine learning approach to grade the video-based PCK assessment. The findings of this study contribute to the field in several ways:

First, the three proposed sources of CIV (Zhai, Haudek, Stuhlsatz, & Wilson, 2020) are not always consistent in their contribution and significance in the two PCK assessment tasks. Though using the same scenarios and teaching video clips, the two tasks were designed to measure different constructs. The first task, which assesses Analyzing Student Thinking, yields significant CIV by the variation of scenarios and the rater sensitivity; while for the second task, targeting Analyzing Responsive Teaching, the variation of scenarios does not yield significant CIV, but the rater severity and the rater sensitivity of scenarios do. This finding supports the idea that the amount of CIV might be associated with the construct of interest. For the Analyzing Student Thinking construct, we recognize that it is easier to identify and describe for the observant teachers, and therefore the teacher average PCK measure is greater than the average item difficulty (i.e., 0), so that the variance of teacher PCK measure is greater. While for the Analyzing Responsive Teaching construct, observant teachers might have more difficulty to identify or use pedagogical language to describe examples of this practice, so the teacher average PCK measure is less than the average item difficulty (i.e., 0). These factors do not only impact the variance of the teacher PCK measure but also the CIV. Therefore, we assert that the CIV should be understood within the scope of the construct of interest, for the sake of better guiding item development.

In a theoretical paper regarding rater cognition, Bejar (2012) declared that the rater "consistency and severity among scorers" really matters and urged researchers to collect evidence. Our study has contributed to this effort by finding that human experts and machine show similar rater severity patterns across two different testing tasks, given the significant difference of both human-human and machine-human agreements between the two constructs respectively. We have also found that the machine is always the most severe rater compared to human raters in both of the assessment tasks in this study. This pattern across the two tasks indicates a very stable conclusion: the severity observed is a feature of the raters, including the machine. The observed patterns for both humans and machine do not change due to the variability of the tasks, nor due to the quality of the algorithmic models. Rater Emerson is the most severe human rater, but the machine is even more severe than Emerson. We presume that this severity of machine scoring might be due to the fact that in order for the machine to assign a positive score to a response, it must collect and identify enough lexical evidence from a response to match the features used as variables in the classification algorithm. On the contrary, humans can often understand the meaning of a response, with more lenient requirements for the exact lexical features of the response. For example, experts can interpret differently worded responses to have the same meaning and thus assign the same score to these different responses, while this may not be true for the machine until it has "learned" all the possible text that indicates that same idea, usually through a larger training data set.

These findings may also relate to the method by which the computer scoring models were generated in this study. For developing the scoring models for each task, responses collected from all three scenarios were used as a single training set of data. This means that the computer scoring model likely used lexical features common in responses across all responses with three scenario contexts and aligned them to the underlying construct. Since the training set had a limited number of teacher responses in each scenario, it is unlikely that infrequent text specific to a given scenario could be "learned" by the machine and incorporated into the scoring model. Therefore, responses that include text that deviates too far from the common features in the model might be scored more severely. Such a limitation would likely not exist for human experts, who understand PCK and the diversity of language to express PCK. Such an issue might be addressed by increasing the number of human-scored responses in the training set of data used to develop the computer model.

This suggests that machine scoring has a limitation compared to human scorers. However, we also note that the machine is not always significantly more severe in assigning scores than human experts. For example, the machine is significantly more severe in assessing the Analyzing Responsive Teaching construct, while it is not significantly different than human raters in assessing the Analyzing Student Thinking construct.

We further found that even if the machine is the most severe rater, it is least sensitive to the scenarios of the two tasks respectively, given the significant difference of both human-human and machine-human agreements between the two constructs respectively. That is, scores assigned by machine have the least variance of sensitivity by the human rater, by scenario, and by tasks, thus minimizing the contributions to the CIV when compared to the three human raters. Please note that this conclusion was made when the variance of scenarios and that of rater severity have already been excluded, according to our model. A counterpart example is the variability of Lacie's ratings on Task I, for which Lacie is the most severe judge for Scenarios 1 and 3 but also the most lenient judge, among the machine and the other two human scorers, for Scenario 2. This finding suggests that human raters might be very sensitive to assessment task scenarios. With such sensitivity, we anticipate Lacie’s scoring would lead to significant CIV in the scores. In contrast, the machine seems the least sensitive to scenarios because it shows the least variation. This finding supports the argument that the machine model might have little repetitive bias because the machine could reliably recognize the critical lexical features of the response without being distracted by the variability of the scenarios or other lexical features present in the response (Clauser et al., 2002). We thus assert that machine scoring might be more reliable in grading measures targeting the same construct with different scenarios and suggest that others monitor CIV in other applications of computer-assisted scoring assessments.

At last, we find it very important to restate the nature of the measurement model, especially toward the CIV from scenarios. We used a measurement model that regards scenario as CIV and found that scenario significantly impacts teachers' scores. We also found that human raters are more sensitive to scenarios than the machine, especially for teachers' analyzing of responsive teaching. This finding confirmed that the scenario would be a source of CIV from the measurement perspective. However, several studies argue that teachers' PCK should be scenario-dependent in nature (Hume, Cooper, & Borowski, 2019). In their view, the variance caused by scenarios might be of interest, rather than CIV. We do not use our findings to dispute this view of the nature of PCK. Rather, we are interested in whether the scores assigned to teachers can support a consistent and valid inference of teachers' PCK. From this perspective, if the variance due to different scenarios is actually construct-of-interest variance, then we cannot use the scores to draw a consistent conclusion about teachers' PCK because this variance is systematically different, rather than randomly. Our suggestion to others who decide to integrate scenario as part of the construct of interest is that their assessment task, rubric, inference, and conclusion about teachers' PCK should consistently include scenarios. We think this shift of approaching PCK as a context-dependent construct will significantly affect the measurement of PCK because many issues emerge and will need further exploration. For example, what factors should we take into consideration when sampling the scenarios for the assessment task for given teachers, or in given grade levels or for given cultures or countries? We suggest that this shift of approaching PCK can be a significant direction for the future measurement of PCK and is consistent with the concern of the grain size for PCK measures in the refined consensus model of PCK (Carlson & Daehler, 2019). In the refined PCK model, contextualized features are highlighted to reflect the context-dependency of PCK. Also, the approach we used to examine CIV for contextualized features of PCK is applicable to examine how teachers’

PCK depends on topics or domains that are addressed in the refined PCK model.

Though the study has contributed to our understanding of machine learning in assessing science teacher PCK, we noted several limitations. First, our study only includes three scenarios and three human raters. Future studies should increase the number of scenarios and human raters to further examine the conclusions. Second, given that we applied quantitative methods, we are not able to explore the mechanisms of some descriptive findings. For example, we did not uncover why the machine was more severe than human raters even though our outcomes indicate this is the case. However, future studies can employ qualitative methods to further explore these reasons.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DGE 1438739 and DUE 1323162. We are grateful to the team members: Drs. Mark Urban-Lurain, John Merrill, Marisol Mercado Santiago, Michael Fleming, Zoë Elizabeth Buck Bracey, and Brian Donovan.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.stueduc.2020.100916.

References

AACR. (2019). Retrieved on April 1st, 2019 from https://apps.beyondmultiplechoice.org.

AERA, & NCME, A. (1999). *Standards for educational and psychological testing. The standards for educational and psychological testing*.

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C. Aggarwal, & C. Zhai (Eds.). *Mining text data*. Boston, MA: Springer.

Alonzo, A., Kobarg, M., & Seidel, T. (2012). Pedagogical content knowledge as reflected in teacher–student interactions: Analysis of two video cases. *Journal of research in science teaching, 49*(10), 1211–1239.

Alonzo, A. C., & Kim, J. (2016). Declarative and dynamic pedagogical content knowledge as elicited through two video-based interview methods. *Journal of Research in Science Teaching, 53*(8), 1259–1286.

Baxter, J. A., & Lederman, N. G. (1999). *Assessment and measurement of pedagogical content knowledge. Examining pedagogical content knowledge*. Springer147–161.

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement Issues and Practice, 31*(3), 2–9.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement Issues and Practice, 30*(1), 3–12.

Carlson, J., & Daehler, K. (2019). *Repositioning of PCK in teachers' professional knowledge: The refined Consensus Model of PCK. Repositioning PCK in teachers' professional knowledge.* Abingdon: Routledge.

Clauser, B., Kane, M., & Swanson, D. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education, 15*(4), 413–432.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

Ebel, R., & Frisbie, D. (1986). Essentials of educational measurement. *Educational Measurement Issues and Practice*.

Gallagher, A., Bennett, R. E., Cahalan, C., & Rock, D. A. (2002). Validity and fairness in technology-based assessment: Detecting construct- irrelevant variance in an open-ended, computerized mathematics task. *Educational Assessment, 8*(1), 27–41.

Gess-Newsome, J., Taylor, J. A., Carlson, J., Gardner, A. L., Wilson, C. D., & Stuhlsatz, M. A. (2019). Teacher pedagogical content knowledge, practice, and student achievement. *International Journal of Science Education, 41*(7), 944–963.

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing, 30*, 21–31.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement Issues and Practice, 23*(1), 17–27.

Hume, A., Cooper, R., & Borowski, A. (2019). *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science.* Singapore: Springer.

Kagan, D. M. (1990). Ways of evaluating teacher cognition: Inferences concerning the Goldilocks principle. *Review of Educational Research, 60*(3), 419–469.

Keller, M. M., Neumann, K., & Fischer, H. E. (2017). The impact of physics teachers' pedagogical content knowledge and motivation on students' achievement and interest. *Journal of Research in Science Teaching, 54*(5), 586–614.

Kirschner, S., Borowski, A., Fischer, H. E., Gess-Newsome, J., & von Aufschnaiter, C. (2016). Developing and evaluating a paper-and-pencil test to assess components of physics teachers' pedagogical content knowledge. *International Journal of Science Education, 38*(8), 1343–1372.

Kirschner, S., Taylor, J., Rollnick, M., Borowski, A., & Mavhunga, E. (2015). Gathering evidence for the validity of PCK measures. In A. Berry, P. Friedrichsen, & J. Loughran (Eds.). *Re-examining pedagogical content knowledge in science education* (pp. 229–241). New York: Routledge.

Kleickmann, T., Richter, D., Kunter, M., Elsner, J., Besser, M., Krauss, S., et al. (2013). Teachers' content knowledge and pedagogical content knowledge: The role of structural differences in teacher education. *Journal of Teacher Education, 64*(1), 90–106.

Large, J., Lines, J., & Bagnall, A. (2019). A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. *Data Mining and Knowledge Discovery, 33*(6), 1674–1709.

Linacre, J. (1989). *Many-facet rasch measurement.* Chicago, IL: University of Chicago Press.

Linacre, J. (2002). The judging debacle in pairs figure skating. *Rasch Measurement Transactions, 15*(4), 839–840.

Liu, X., & McKeough, A. (2005). Developmental growth in students' concept of energy: Analysis of selected items from the TIMSS database. *Journal of Research in Science Teaching, 42*(5), 493–517.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014a). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement Issues and Practice, 33*(2), 19–28.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014b). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement Issues and Practice, 33*(2), 19–28.

Magnusson, S., Krajcik, J., & Borko, H. (1999). *Nature, sources, and development of pedagogical content knowledge for science teaching. Examining pedagogical content knowledge*. Springer95–132.

McNeill, K. L., González-Howard, M., Katsh-Singer, R., & Loper, S. (2016). Pedagogical content knowledge of argumentation: Using classroom contexts to assess high-quality PCK rather than pseudo argumentation. *Journal of Research in Science Teaching, 53*(2), 261–290.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21*, 215–237.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology, 21*(1), 183–196.

Ooi, P., & Engelhard, J. G. (2019). Examining rater judgements in music performance assessment using many-facets rasch rating scale measurement model. *Journal of Applied Measurement, 20*(1), 79–99.

Park, S., & Oliver, J. S. (2008). Revisiting the conceptualization of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education, 38*(3), 261–284.

Santagata, R. (2009). Designing video-based professional development for mathematics teachers in low-performing schools. *Journal of Teacher Education, 60*(1), 38–51.

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment, 20*(1), 46–65.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Wilson, C. D., Borowski, A., & van Driel, J. (2019). *Perspectives on the future of PCK research in science education and beyond. Repositioning pedagogical content knowledge in teachers' knowledge for teaching science.* Springer289–300.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Rasch measurement*.

Wu, M. L., Adams, R., Wilson, M., & Haldane, S. (2007). *ACER conquest version 2.0.* Camberwell, Victoria, Australia: ACER Press, Australian Council for Educational Research.

Zaichkowsky, J. L. (1985). Measuring the involvement construct. *The Journal of Consumer Research, 12*(3), 341–352.

Zehner, F., Saelzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement, 76*(2), 280–303.

Zhai, X. (2019). Applying machine learning in science assessment: Opportunity and challenge. *Journal of Science Education and Technology,* 1–4. https://doi.org/10.13140/RG.2.2.10914.07365.

Zhai, X., Ruiz-Primo, M. A., Li, M., Kanopka, K., Hernandez, P., Dong, D., & Minstrell, J. (2019). *Students' involvement in contextualized science assessment.* Baltimore, MD: National Association of Research in Science Teaching.

Zhai, X., Haudek, K., Shi, L., Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching,* 1–30. https://doi.org/10.1002/tea.21658 In press.

Zhai, X., Haudek, K., Stuhlsatz, M., & Wilson, C. (2020). Examining construct-irrelevant variances of contextualized constructed-response assessment: A many-facet Rasch modeling approach. *Disciplinary and Interdisciplinary Science Education Research* Submitted for publication.

Zhai, X., Krajcik, J., & Pellegrino, J. (2020). On the validity of machine learning-based Next Generation Science Assessments: An inferential validity network. *Journal of Science Education and Technology* Submitted for publication.

Zhai, X., Yin, Y., Pellegrino, J., Haudek, K., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757.

Zhang, M., Lundeberg, M., Koehler, M. J., & Eberhardt, J. (2011). Understanding affordances and challenges of three types of video for teacher professional development. *Teaching and Teacher Education, 27*(2), 454–462.

Xiaoming Zhai is an Assistant Professor in Science Education. He is interested in developing and using innovative assessment and technology to support science teaching and learning.

Kevin C. Haudek is an Assistant Professor. He is interested in the application of computerized tools to evaluate student writing, content-rich responses in science education.

Molly Stuhlsatz is a Research Scientist. She is interested in assessment of teacher and student learning in science education.

Christopher Wilson is a Senior Research Scientist and Director of Research at BSCS Science Learning. His research focuses on the assessment of teacher and student learning in science education, and the application of automated scoring techniques to teacher PCK.